

Geometric ergodicity of the Random Walk Metropolis with position-dependent proposal covariance.

Samuel Livingstone

*Department of Statistical Science, University College London,
Gower Street, London WC1E 6BT, United Kingdom.*

Abstract

We consider a Metropolis–Hastings method with proposal kernel $\mathcal{N}(x, hG^{-1}(x))$, where x is the current state. After discussing specific cases from the literature, we analyse the ergodicity properties of the resulting Markov chains. In one dimension we find that suitable choice of $G^{-1}(x)$ can change the ergodicity properties compared to the Random Walk Metropolis case $\mathcal{N}(x, h\Sigma)$, either for the better or worse. In higher dimensions we use a specific example to show that judicious choice of $G^{-1}(x)$ can produce a chain which will converge at a geometric rate to its limiting distribution when probability concentrates on an ever narrower ridge as $|x|$ grows, something which is not true for the Random Walk Metropolis.

Keywords: Monte Carlo, MCMC, Markov chains, Computational Statistics, Bayesian Inference.

1. Introduction

Markov chain Monte Carlo (MCMC) methods are techniques for estimating expectations with respect to some distribution $\pi(\cdot)$, which need not be normalised. This is done by sampling a Markov chain which has limiting distribution $\pi(\cdot)$, and computing empirical averages. A popular form of MCMC is the Metropolis–Hastings algorithm [1, 2], where at each time step a ‘proposed’ move is drawn from some candidate distribution, and then accepted with some probability, otherwise the chain stays at the current point. Interest lies in finding choices of candidate distribution that will produce sensible estimators for expectations with respect to $\pi(\cdot)$.

The quality of these estimators can be assessed in many different ways, but a common approach is to understand conditions on $\pi(\cdot)$ that will result in a chain which converges to its limiting distribution at a *geometric* rate. If such a rate can be established, then a Central Limit Theorem will exist for expectations

Email address: samuel.livingstone@ucl.ac.uk (Samuel Livingstone)

of functionals with finite second absolute moment under $\pi(\cdot)$ if the chain is reversible.¹

A simple yet often effective choice is a symmetric candidate distribution centred at the current point in the chain (with a fixed variance), resulting in the *Random Walk Metropolis* (RWM) (e.g. [3]). The convergence properties of a chain produced by the RWM are well-studied. In one dimension, essentially convergence is geometric if $\pi(x)$ decays at an exponential or faster rate in the tails [4], while in higher dimensions an additional curvature condition is required [5]. Slower rates of convergence have also been established in the case of heavier tails [6].

Recently, some MCMC methods have been proposed which generalise the RWM, whereby proposals are still centred at the current point x and symmetric, but the variance changes with x [7, 8, 9, 10, 11]. An extension to infinite-dimensional Hilbert spaces is also suggested in [12]. The motivation is that the chain can become more ‘local’, perhaps making larger jumps when out in the tails, or mimicking the local dependence structure of $\pi(\cdot)$ to propose more intelligent moves. Designing MCMC methods of this nature is particularly relevant for modern Bayesian inference problems, where posterior distributions are often high dimensional and exhibit nonlinear correlations [13]. We term this approach the *Position-Dependent Random Walk Metropolis* (PDRWM), although technically this is a misnomer, since proposals are no longer random walks.² Other choices of candidate distribution designed with distributions that exhibit nonlinear correlations were introduced in [13]. Although powerful, these require derivative information for $\log \pi(x)$, something which can be unavailable in modern inference problems (e.g. [14]). We note that no such information is required for the PDRWM, as evidenced by the particular cases suggested in [7, 8, 9, 10, 11]. However, there are relations between the approaches, to the extent that understanding how the properties of the PDRWM differ from the standard RWM should also aid understanding of the methods introduced in [13].

In this article we consider the convergence rate of a Markov chain generated by the PDRWM to its limiting distribution. Our main interest lies in whether this generalisation can change these *ergodicity* properties compared to the standard RWM with fixed covariance. We focus on the case where the candidate distribution is Gaussian, and in one dimension we establish necessary and sufficient growth conditions on the proposal variance and tail behaviour of $\pi(x)$ for geometric ergodicity. Some of the results extend naturally to higher dimensions, but we also offer an illustrative example showing that the curvature condition can be alleviated when the proposal covariance is allowed to change with position. In Section 2 necessary concepts about Markov chains are briefly reviewed, before the PDRWM is introduced in Section 3. One dimensional results are given in Section 4, before those for higher dimensions in Section 5 and a discus-

¹We deal exclusively with reversible chains here, in the non-reversible case the requirement is a finite $(2 + \delta)$ th absolute moment.

²The size of jump now depends on the current position in the chain.

sion in Section 6. Throughout $\pi(\cdot)$ denotes a probability distribution, and $\pi(x)$ its density with respect to Lebesgue measure.

2. Markov Chains & Geometric Ergodicity

We will work on the measurable space $(\mathcal{X}, \mathcal{B})$, so that each $X_t \in \mathcal{X}$ for a discrete-time Markov chain $\{X_t\}_{t \geq 0}$ with time-homogeneous transition kernel $P : \mathcal{X} \times \mathcal{B} \rightarrow [0, 1]$, where $P(x, A) = \mathbb{P}[X_{i+1} \in A | X_i = x]$ and $P^n(x, A)$ is defined similarly for X_{i+n} . All chains we consider will have invariant distribution $\pi(\cdot)$, and be both π -irreducible and aperiodic, meaning $\pi(\cdot)$ is the limiting distribution from π -almost any starting point [15]. We use $|\cdot|$ to denote the Euclidean norm.

In Markov chain Monte Carlo the objective is to construct estimators of $\mathbb{E}_\pi[f]$, for some $f : \mathcal{X} \rightarrow \mathbb{R}$, by computing

$$\hat{f}_n = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad X_i \sim P^i(x_0, \cdot).$$

If $\pi(\cdot)$ is the limiting distribution for the chain then P will be *ergodic*, meaning $\hat{f}_n \xrightarrow{a.s.} \mathbb{E}_\pi[f]$ from π -almost any starting point. For finite n the quality of \hat{f}_n intuitively depends on how quickly $P^n(x, \cdot)$ approaches $\pi(\cdot)$. We call the chain *geometrically ergodic* if

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq M(x)\rho^n, \quad (1)$$

from π -almost any $x \in \mathcal{X}$, for some $M > 0$ and $\rho < 1$, where $\|\mu(\cdot) - \nu(\cdot)\|_{TV} := \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|$ is the total variation distance between distributions $\mu(\cdot)$ and $\nu(\cdot)$ [15].

Geometric ergodicity implies that if $\mathbb{E}_\pi[|f|^{2+\delta}] < \infty$ for some $\delta > 0$, then

$$\sqrt{n} \left(\hat{f}_n - \mathbb{E}_\pi[f] \right) \xrightarrow{d} \mathcal{N}(0, v(P, f)), \quad (2)$$

for some asymptotic variance $v(P, f)$. Equation (2) enables the construction of asymptotic confidence intervals for \hat{f}_n [15]. Several techniques now exist for constructing *non-asymptotic* confidence intervals (e.g. [16, 17, 18]), but at present it is not yet clear whether these can be applied in the same sort of generality as (2). In some cases, such approaches rely on either geometric ergodicity or the equivalent³ condition of a *spectral gap* existing for P [19], so (1) must also be established for many of these non-asymptotic results to hold (e.g. [17]). Geometric ergodicity is also often a requirement in establishing the stability of *noisy* Markov chains in which P is approximated due to either intractability or computational convenience [20, 21] (in other instances slightly weaker but related conditions are needed [22]).

³This is true for reversible chains.

In practice, geometric ergodicity does not guarantee that \hat{f}_n will be a sensible estimator, as $M(x)$ can be arbitrarily large if the chain is initialised far from the typical set under $\pi(\cdot)$, and ρ may be very close to 1. However, chains which are not geometrically ergodic can often either get ‘stuck’ for a long time in low-probability regions or fail to explore the entire distribution adequately, sometimes in ways which are difficult to diagnose using standard MCMC diagnostics.

2.1. Establishing geometric ergodicity

It is shown in Chapter 15 of [23] that (1) is equivalent to the condition that there exists a *Lyapunov* function $V : \mathcal{X} \rightarrow [1, \infty)$ and some $\lambda < 1, b < \infty$ such that

$$PV(x) \leq \lambda V(x) + b \mathbb{1}_C(x), \quad (3)$$

where $PV(x) := \int V(y)P(x, dy)$. The set $C \subset \mathcal{X}$ must be *small*, meaning that for some $m \in \mathbb{N}$, $\varepsilon > 0$ and probability measure $\nu(\cdot)$

$$P^m(x, A) \geq \varepsilon \nu(A), \quad (4)$$

for any $x \in C$ and $A \in \mathcal{B}$. Equations (3) and (4) are referred to as *drift* and *minorisation* conditions. Intuitively, C can be thought of as the centre of the space, and (3) ensures that some one dimensional projection of $\{X_t\}_{t \geq 0}$ drifts towards C at a geometric rate when outside. In fact, (3) is sufficient for the return time distribution to C to have geometric tails [23]. Once in C , (4) ensures that with some probability the chain forgets its past and hence *regenerates*. This regeneration allows the chain to ‘couple’ with another started at stationarity, giving a bound on the total variation distance through the *coupling inequality* [15]. More intuition is given in [24].

Transition kernels considered here will be of the *Metropolis–Hastings* type, given by

$$P(x, dy) = \alpha(x, y)Q(x, dy) + r(x)\delta_x(dy), \quad (5)$$

where $Q(x, dy) = q(y|x)dy$ is some candidate kernel, α is the ‘acceptance rate’ and $r(x) = 1 - \int \alpha(x, y)Q(x, dy)$. Here we choose

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, \quad (6)$$

where $a \wedge b$ denotes the minimum of a and b . This choice implies that P satisfies detailed balance for $\pi(\cdot)$ [25], and hence the chain is reversible (note that other choices for α can result in non-reversible chains, see [26] for details). In this case (2) applies to a slightly broader class of functionals, namely those with $\mathbb{E}_\pi[|f|^2] < \infty$ [19].

Roberts & Tweedie [5], following on from [23], introduced the following regularity conditions.

Theorem 1. (*Roberts & Tweedie*). Suppose that $\pi(x)$ is bounded away from 0 and ∞ on compact sets, and there exists $\delta_q > 0$ and $\varepsilon_q > 0$ such that, for every x

$$|x - y| \leq \delta_q \Rightarrow q(y|x) \geq \varepsilon_q.$$

Then the chain with kernel (5) is μ^{Leb} -irreducible and aperiodic, and every nonempty compact set is small.

For the choices of Q considered in this article these conditions hold, and we will restrict ourselves to forms of $\pi(x)$ for which the same is true (apart from a specific case in Section 5). Under Theorem 1 then (1) only holds if a Lyapunov function $V : \mathcal{X} \rightarrow [1, \infty]$ with $\mathbb{E}_\pi[V] < \infty$ exists such that

$$\limsup_{|x| \rightarrow \infty} \frac{PV(x)}{V(x)} < 1. \quad (7)$$

When P is of the Metropolis-Hastings type, (7) can be written

$$\limsup_{|x| \rightarrow \infty} \int \left[\frac{V(y)}{V(x)} - 1 \right] \alpha(x, y) Q(x, dy) < 0. \quad (8)$$

In this case a simple criterion for lack of geometric ergodicity is

$$\limsup_{|x| \rightarrow \infty} r(x) = 1. \quad (9)$$

Intuitively this implies that the chain is likely to get ‘stuck’ in the tails of a distribution for large periods.

Jarner & Tweedie [27] introduce a necessary condition for geometric ergodicity through a *tightness* condition.

Theorem 2. (*Jarner & Tweedie*). If for any $\varepsilon > 0$ there is a $\delta > 0$ such that for all $x \in \mathcal{X}$

$$P(x, B_\delta(x)) > 1 - \varepsilon,$$

where $B_\delta(x) := \{y \in \mathcal{X} : d(x, y) < \delta\}$, then P can only be geometrically ergodic if for some $s > 0$

$$\int e^{s|x|} \pi(dx) < \infty.$$

The result highlights that when $\pi(\cdot)$ is heavy-tailed the chain must be able to make very large moves and still be capable of returning to the centre quickly for (1) to hold. In the Metropolis-Hastings case it is straightforward to see that

$$Q(x, B_\delta(x)) > 1 - \varepsilon \Rightarrow P(x, B_\delta(x)) > 1 - \varepsilon,$$

which is a useful approach to establishing lack of (1) in the heavy-tailed case.

3. Position-dependent Random Walk Metropolis

In the RWM, $Q(x, dy) = q(|y - x|)dy$, meaning (6) reduces to $\alpha(x, y) = 1 \wedge \pi(y)/\pi(x)$. A common choice is $Q(x, \cdot) = \mathcal{N}(x, h\Sigma)$, with Σ chosen to mimic the global covariance structure of $\pi(\cdot)$ [3]. Various results exist concerning the optimal choice of h in a given setting (e.g. [28]). It is straightforward to see that Theorem 2 holds here, so that the tails of $\pi(x)$ must be uniformly exponential or lighter for geometric ergodicity. In one dimension this is in fact a sufficient condition [4], while for higher dimensions additional conditions are required [5]. We return to this case in Section 5.

In the PDRWM $Q(x, \cdot) = \mathcal{N}(x, hG^{-1}(x))$, so (6) becomes

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)|G(y)|^{\frac{1}{2}}}{\pi(x)|G(x)|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - y)^T[G(y) - G(x)](x - y)\right).$$

The intuition here is that proposals are more able to reflect the local dependence structure of $\pi(\cdot)$. In some cases this dependence may vary greatly in different parts of the state-space, making a global choice of Σ ineffective [9].

Readers familiar with differential geometry will recognise the volume element $|G(x)|^{1/2}dx$ and the linear approximations to the distance between x and y taken at each point through $G(x)$ and $G(y)$ if \mathcal{X} is viewed as a Riemannian manifold with metric G . We do not explore these observations further here, but the interested reader is referred to [29] for more discussion.

The choice of $G(x)$ is an obvious question. In fact, specific variants of this method have appeared on many occasions in the literature, some of which we now summarise.

1. *Tempered Langevin diffusions* [8] $G^{-1}(x) = \pi^{-1}(x)I$. The authors highlight that the diffusion with dynamics $dX_t = \pi^{-\frac{1}{2}}(X_t)dW_t$ has invariant distribution $\pi(\cdot)$, motivating the choice. The method was shown to perform well for a bi-modal $\pi(x)$, as larger jumps are proposed in the low density region between the two modes.
2. *State-dependent Metropolis* [7] $G^{-1}(x) = (1 + |x|)^b$. Here the intuition is simply that $b > 0$ means larger jumps will be made in the tails. In one dimension the authors compare the expected squared jumping distance $\mathbb{E}[(X_{i+1} - X_i)^2]$ empirically for chains exploring a $\mathcal{N}(0, 1)$ target distribution, choosing b adaptively, and found $b \approx 1.6$ to be optimal.
3. *Regional adaptive Metropolis-Hastings* [7, 11]. $G^{-1}(x) = \sum_{i=1}^m \mathbb{1}_{x \in \mathcal{X}_i} \Sigma_i$. In this case the state-space is partitioned into $\mathcal{X}_1 \cup \dots \cup \mathcal{X}_m$, and a different proposal covariance Σ_i is learned adaptively in each region $1 \leq i \leq m$. An extension which allows for some errors in choosing an appropriate partition is discussed in [11].
4. *Localised Random Walk Metropolis* [10]. $G^{-1}(x) = \sum_{k=1}^m \tilde{q}_\theta(k|x) \Sigma_k$. Here $\tilde{q}_\theta(k|x)$ are weights based on approximating $\pi(x)$ with some mixture of Normal/Student's t distributions, using the approach suggested in [30]. At each iteration of the algorithm a mixture component k is sampled from $\tilde{q}_\theta(\cdot|x)$, and the covariance Σ_k is used for the proposal $Q(x, dy)$.

5. *Kernel adaptive Metropolis–Hastings* [9]. $G^{-1}(x) = \gamma^2 I + \nu^2 M_x H M_x^T$, where $M_x = 2[\nabla_x k(z_1, x), \dots, \nabla_x k(z_n, x)]$ for some kernel function k and n past samples $\{z_1, \dots, z_n\}$, $H = I - 1/n \mathbb{1}_{n \times n}$ is a centering matrix, and γ, ν are tuning parameters. The approach is based around performing nonlinear principal components analysis on past samples from the chain to learn a local covariance. Illustrative examples for the case of a Gaussian kernel show that $M_x H M_x^T$ acts as a weighted empirical covariance of samples z , with larger weights given to the z_i which are closer to x [9].

The latter cases also motivate any choice of the form

$$G^{-1}(x) = \sum_{i=1}^n w(x, z_i) (z_i - x)^T (z_i - x)$$

for some past samples $\{z_1, \dots, z_n\}$ and weight function $w : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ with $\sum_i w(x, z_i) = 1$ that decays as $|x - z_i|$ grows, which would also mimic the local curvature of $\pi(\cdot)$ (taking care to appropriately regularise and diminish adaptation so as to preserve ergodicity, as outlined in [10]). The logic of [13, 31] could also be applied, by choosing $G(x)$ as some regularised version of the negative Hessian of $\log \pi(x)$. However, if such derivative information were available it would seem more sensible to use a more sophisticated method than a martingale proposal (see e.g. [13]).

4. Results in One Dimension

Here the specific choice of $G(x)$ is left open, and we instead consider two different general scenarios as $|x| \rightarrow \infty$, i) $G^{-1}(x) \rightarrow \Sigma$, and ii) $G^{-1}(x) \rightarrow \infty$ at some rate. In theory there is also the possibility that $G^{-1}(x) \rightarrow 0$, though intuitively this would not seem to be a particularly sensible choice as chains would be extremely likely to spend a long time in the tails of a distribution, so we do not consider it.

Three scenarios are considered for the tail behaviour of $\pi(x)$. We refer to this density as *log-concave in the tails* if for some $x_0 > 0$ and $a > 0$

$$\pi(y)/\pi(x) \leq e^{-a(y-x)}, \quad \forall y \geq x \geq x_0, \quad (10)$$

and a similar condition holds in the negative tail. If (10) is not satisfied but there is some $\beta \in (0, 1)$ such that the above condition can be replaced with $\pi(y)/\pi(x) \leq \exp\{-a(y^\beta - x^\beta)\}$, then we call the density subexponential (note this is not the standard definition). Finally, we call $\pi(x)$ ‘polynomial-tailed’ if $\pi(x) \propto |x|^{-p}$ for large $|x|$ and some $p \geq 1$. We also apply asymptotic growth conditions for $G^{-1}(x)$, and without loss of generality assume that these hold for any x larger than the same x_0 in absolute value.

We introduce some asymptotic notation in this section. For positive real-valued functions f and g , let $f(x) = \Theta(g(x))$ imply $f(x)/g(x) \rightarrow C > 0$ as $x \rightarrow \infty$, and $f(x) = \omega(g(x))$ imply $f(x)/g(x) \rightarrow \infty$. The more familiar big-O

and little-o notation is also used. The main results of this section are summarised in Table 1 at the end of the section.

The first result emphasises a growing variance as a necessary requirement for geometric ergodicity in the heavy-tailed case.

Lemma 1. *If $G^{-1}(x) \leq \sigma^2$, then the PDRWM can only produce a geometrically ergodic Markov chain if $\pi(x)$ is log-concave in the tails.*

Proof: In this case for any choice of $\varepsilon > 0$ there is a $\delta > 0$ such that $Q(x, B_\delta(x)) > 1 - \varepsilon$, so Theorem 2 can be applied. ■

Though the heavy-tailed case is a challenging scenario, the standard RWM with fixed covariance will produce a geometrically ergodic Markov chain if $\pi(x)$ is log-concave. Next we extend this result to the case of sub-quadratic variance growth in the tails.

Lemma 2. *If $G^{-1}(x) = o(|x|^2)$ and $\pi(x)$ is log-concave in the tails, then the PDRWM method produces a geometrically ergodic Markov chain from π -almost any starting point. If $\pi(x)$ is subexponential for some $\beta \in (0, 1)$, then choosing $G^{-1}(x) = \Theta(|x|^\gamma)$ for some $2(1 - \beta) < \gamma < 2$ gives the same result.*

Proof: See Appendix A.1.

The log-concave proof consists of partitioning \mathcal{X} into five regions, and showing that as $|x| \rightarrow \infty$, (8) evaluated over each of these regions will either become arbitrarily small or remain strictly negative. We use the Lyapunov function $V(x) = e^{s|x|}$ for some $s > 0$. This choice allows results about moment generating functions of truncated Gaussian distributions (see Appendix B) to be used, in conjunction with simple bounds on the cumulative distribution function from [32], to establish that (8) will become arbitrarily small for regions of \mathcal{X} outside the ‘typical set’ $(x - cx^{\gamma/2}, x + cx^{\gamma/2})$. Theorem 3.2 from [4] shows that for the RWM with fixed covariance (8) evaluated over this region will be strictly negative. The essence of the argument is that for $y > x$ in the tails, $\alpha_R(x, y) \leq e^{-a(y-x)}$ by log-concavity, so as long as s is chosen to be less than a this decay will dominate any growth in $V(y)$ here. As for any inwards proposals $\alpha_R(x, y) = 1$ then it can be shown that (8) is strictly negative when evaluated over this region.

The crucial additional difficulty in the case of growing covariance is that the acceptance rate in this region (for suitably large x) is now

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \exp \left(\frac{\gamma}{2} \log \left| \frac{x}{y} \right| - \frac{1}{2h} \left[\frac{(x-y)^2}{y^\gamma} - \frac{(x-y)^2}{x^\gamma} \right] \right)$$

The problematic term lies inside the square bracket: this will be negative for $y > x$, meaning a large positive component in $\alpha(x, y)$. To deal with this, we

use a Taylor expansion of $y^{-\gamma}$ about x and some simplifications to show that provided $\gamma < 2$, for large enough x , *locally* (for y near x , where the choice of region plays a role) the acceptance rate will still satisfy

$$\alpha(x, y) = 1 \text{ for } y < x, \quad \alpha(x, y) \leq e^{-a(y-x)+\delta_x}, \text{ for } y > x,$$

where δ_x can be made arbitrarily small. This allows us to use a similar argument to that in [4] to prove the result. Outside of this region the Gaussian tails of $Q(x, \cdot)$ take care of any less desirable behaviour of $\alpha(x, y)$. To extend this result to the subexponential case, we choose $V(x) = e^{s|x|^\beta}$, and Taylor expand $|y|^\beta$ in the typical set to get a suitable bound on $\alpha(x, y)$.

Note that this lemma includes as a special case any instance in which $G^{-1}(x) \uparrow \sigma^2$ as $|x| \rightarrow \infty$. However, the case $G^{-1}(x) \rightarrow \sigma^2$ from any direction is actually more straightforward to show, by simply moving x far enough into the tails that $G^{-1}(x) \approx \sigma^2$ for all $y \in (x - cx^{\gamma/2}, x + cx^{\gamma/2})$. In this case the argument in [4] can be applied more straightforwardly.

Although we do not formally prove that the method will not produce a geometrically ergodic chain in the polynomial tailed case when $G^{-1}(x) = o(|x|^2)$, we show intuitively that this will be the case. Assuming that in the tails $\pi(x) \propto |x|^{-p}$ for some $p > 1$ then for large x

$$\alpha(x, x + cx^{\gamma/2}) = 1 \wedge \left(\frac{x}{x + cx^{\gamma/2}} \right)^{p+\gamma/2} \exp \left(-\frac{c^2 x^\gamma}{2h} \left[\frac{1}{(x + cx^{\gamma/2})^\gamma} - \frac{1}{x^\gamma} \right] \right).$$

The first expression on the right hand side converges to 1 as $x \rightarrow \infty$, which is akin to the case of fixed proposal covariance. The second term will be larger than one for $c > 0$ and less than one for $c < 0$. So the algorithm will exhibit the same ‘random walk in the tails’ behaviour which is often characteristic of the RWM in this scenario, and so the acceptance rate will fail to enforce a geometric drift back into the centre of the space.

In the case where $\gamma = 2$ this will not happen, as the terms in the above expression will be roughly constant with x . We examine this case next.

Lemma 3. *If $G^{-1}(x) = \Theta(|x|^2)$, then there is a $h_0 > 0$ such that for a step-size $h \in (0, h_0)$ the PDRWM method produces a geometrically ergodic Markov chain from π -almost any starting point, provided $\pi(x) \leq |x|^{-p}$ in the tails for some $p > 1$.*

Proof: See Appendix A.2.

Here the intuition is that proposals in the tails will take the form $y = (1 + \xi\sqrt{h})x$, which if h is chosen to be small will be similar to $y = e^{\xi\sqrt{h}}x$. The latter scheme is sometimes called the *multiplicative* RWM, and is known to be geometrically ergodic in this scenario (e.g. [3]), as this equates to taking a log-transformation of x , which ‘lightens’ the tails of the target density to the point where it becomes log-concave.

In this case we take the Lyapunov function $V(x) = 1 \vee |x|^s$, with $s > 0$ chosen such that $\int V(y)\pi(dy) < \infty$. We again divide the integral (8) into regions, but in this case we show that each of these can be appropriately bounded simply as functions of the step-size h , i.e. independently of x . By examining each term, we show that for a small enough h the integral will be strictly negative.

The result is positive, but in this case is perhaps an example where the theory does not necessarily translate into an effective scheme in practice. If $\pi(x)$ has particularly heavy tails, for example, then it is likely that an extremely small value of h would be needed to ensure (1), meaning the geometric rate of convergence ρ would be close to one. Nonetheless, it is an example of how appropriate choice of $G^{-1}(x)$ can *favourably* change the ergodicity properties of a sampler.

The final result of this section provides a note of warning, that lack of care in choosing $G^{-1}(x)$ can have severe consequences for the method.

Lemma 4. *If $G^{-1}(x) = \omega(|x|^2)$, then the PDRWM method can never produce a geometrically ergodic Markov chain provided $\pi(x) \rightarrow 0$ as $|x| \rightarrow \infty$.*

Proof: See Appendix A.3.

The intuition for this result is straightforward when explained. In the tails, the average proposals will be of size $|x|^{\gamma/2}$, which will be much larger than $|x|$ if $\gamma > 2$, meaning most will send the chain even further into the tails in either direction (and hence will likely be rejected). To make this rigorous we show that (9) holds here, by considering the set of proposals $A_{x,\epsilon} := \{y \in \mathcal{X} : \alpha(x, y) \geq \epsilon\}$, and showing that $Q(x, A_{x,\epsilon}) \rightarrow 0$ as $|x| \rightarrow \infty$, for any $\epsilon > 0$. A specific example is illustrated in Figure 1.

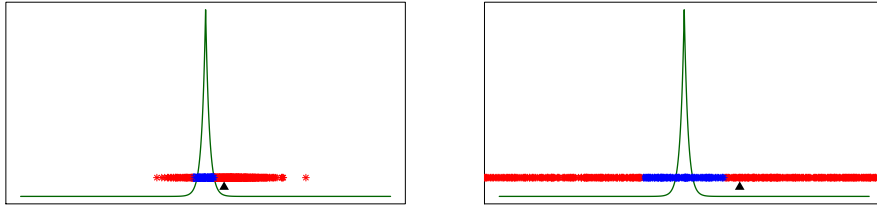


Figure 1: Example with $\pi(x) \propto e^{-|x|}$, $G^{-1}(x) \propto |x|^4$. The black triangle denotes the current state, points highlighted in blue represent proposals with $\alpha(x, y) > 0.5$, with all others highlighted in red. For large $|x|$ the majority of proposals miss the centre of the space and are rejected.

The main results of this section are summarised in Table 1.

Variance	Polynomial Tails	Subexponential	Log-concave
$G^{-1}(x) = o(x ^2)$	\times	\checkmark^+	\checkmark
$G^{-1}(x) = \Theta(x ^2)$	\checkmark^*	\checkmark^*	\checkmark^*
$G^{-1}(x) = \omega(x ^2)$	\times	\times	\times

Table 1: Summary of one dimensional results. Here $f(x) = \omega(g(x))$ means $f/g \rightarrow \infty$ as $x \rightarrow \infty$, $f(x) = \Theta(g(x))$ means $f/g \rightarrow C > 0$, \checkmark means geometrically ergodic, \checkmark^+ means geometrically ergodic provided $G^{-1}(x) \in \Theta(|x|^\gamma)$ for some $2 > \gamma > 2(1 - \beta)$, and \checkmark^* means geometrically ergodic provided h is suitably small.

5. Higher Dimensions

Some results from the previous section naturally carry over to higher dimensions. The most straightforward is outlined below.

Lemma 5. *If each element of $G^{-1}(x)$ is bounded above (uniformly in x), then the PDRWM can only produce a geometrically ergodic Markov chain if the tails of $\pi(x)$ are uniformly exponential or lighter.*

Proof: As with Lemma 1, a straightforward application of Theorem 2 gives the result. \blacksquare

It is also intuitive that an analogue to Lemma 4 will exist here. Specifically, if any diagonal component of the covariance $G^{-1}(x)$ grows at a faster than quadratic rate with x , then the sampler is likely to run into the same difficulties in the tails. Similarly, when $G^{-1}(x) \rightarrow \Sigma$, it is straightforward to see that the sampler will inherit the geometric ergodicity properties of the RWM with fixed covariance, by a similar argument to that discussed for the proof of Lemma 2 in this case.

As mentioned earlier, in the case $G^{-1}(x) = \Sigma$, additional conditions on $\pi(x)$ are required for geometric ergodicity in more than one dimension, outlined in [5]. An example is also given in the paper of the simple two-dimensional density $\pi(x, y) \propto \exp(-x^2 - y^2 - x^2 y^2)$, which fails to meet this criterion. The difficult models are those for which probability concentrates on a ‘ridge’ in the tails, which becomes ever narrower as $|x|$ increases. In this instance, proposals from the RWM are less and less likely to be accepted as $|x|$ grows. The problem is illustrated graphically in Figure 2. Such densities are often encountered as posterior distributions in hierarchical models, with another well-known example being the ‘funnel’, discussed in [33]. On the same figure there is some graphical evidence that if the proposal covariance is allowed to adjust then this problem can be alleviated somewhat.

To explore this more concretely, we design an extremely simple two dimensional density which exhibits the same features, which we call the ‘rectangle’

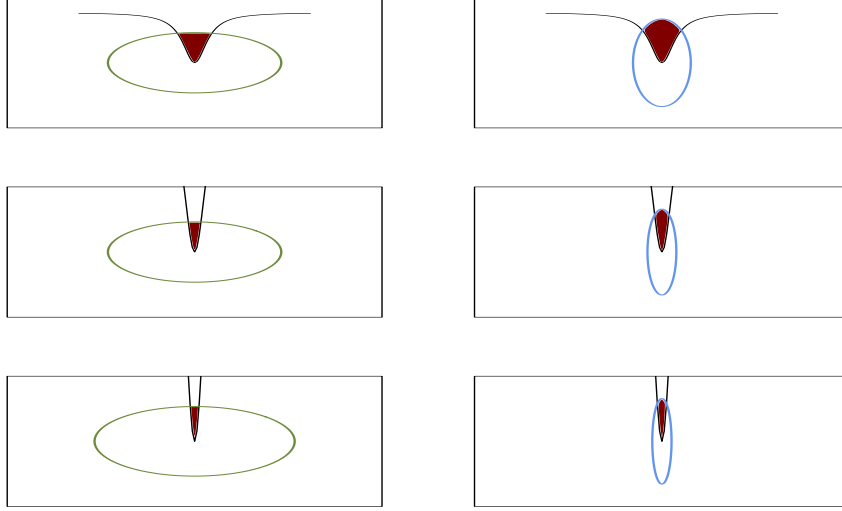


Figure 2: Contours of the density $\pi(x, y) \propto \exp(-x^2 - y^2 - x^2 y^2)$. The left-hand plots show that a RWM with spherical covariance will find it increasingly difficult to propose values which will be accepted as the chain moves into the tails. The right-hand plots suggest that allowing the covariance to change with position might alleviate this issue.

density

$$\square(x) \propto 3^{-\text{int}(x_2)} \mathbb{1}_R(x), \quad R := \{y \in \mathbb{R}^2; y_2 \geq 1, |y_1| \leq 3^{1-\text{int}(y_2)}\},$$

where $\text{int}(z)$ is the integer part of $z \in \mathbb{R}$. This is simply a distribution defined over a sequence of rectangles on the upper-half plane on \mathbb{R}^2 (starting at $y_2 = 1$), each centred on the vertical axis, with height one and with each successive triangle a third of the width and depth of the previous. Intuitively, the density is an ever narrowing staircase, as shown in Figure 3.

For simplicity here we take the Random Walk Metropolis proposal as simply a uniform distribution on the circle of radius one about the current point, so $Q_R(x, A) = |A \cap S_x|/|S_x|$, where $S_x := \{y \in \mathbb{R}^2; |y - x| \leq 1\}$. To imitate the changing covariance in the PDRWM, we take as a proposal a uniform distribution over an ellipse for which the width is $3^{1-\text{int}(x_2)}$ if the current position is $x = (x_1, x_2) \in \mathbb{R}^2$, so $Q_P(x, A) = |A \cap E_x|/|E_x|$, where $E_x = \{y \in \mathbb{R}^2 : 3^{2(1-\text{int}(x_2))}(y_1 - x_1)^2 + (y_2 - x_2)^2 \leq 1\}$. For these choices many of the calculations required in this section reduce to calculating areas of rectangles and ellipses.

The rectangle density does not satisfy the conditions of Theorem 1, as $\square(x)$ is not bounded away from zero on compact sets, however any small set here must still be compact for both methods specified. To see this, note that for any fixed $m < \infty$, $\text{supp}\{P_R^m(x, \cdot)\}$ is compact, so that for a minorisation condition of the form (4) to hold within some small set C , then we must have that $\text{supp}\{\nu(\cdot)\} \subset$

$\text{supp}\{P_R^m(x, \cdot)\} \cap \text{supp}\{P_R^m(y, \cdot)\}$ for every $x, y \in C$. As this intersection will only be non-empty for bounded $|x - y|$, C must be compact. The same argument holds for the elliptical case. Because of this, establishing (9) is still sufficient to characterise lack of geometric ergodicity.

Lemma 6. *The Metropolis–Hastings algorithm with proposal Q_R does not produce a geometrically ergodic Markov chain when $\pi(x) = \square(x)$.*

Proof: It is sufficient to construct a sequence of points $x_p \in \mathbb{R}^2$ such that $|x_p| \rightarrow \infty$ as $p \rightarrow \infty$, and show that $r(x_p) \rightarrow 1$. Take $x_p = (0, p)$ for $p \in \mathbb{N}$. In this case $r(x_p)$ is bounded below by one minus the area of the rectangles that x_p is on the boundary of divided by the area of the circle $|S_x| = \pi$. So we have

$$r(x_p) \geq 1 - \left(\frac{1}{3^{p-2}\pi} + \frac{1}{3^{p-1}\pi} \right) \rightarrow 1$$

as $p \rightarrow \infty$, as required. \blacksquare

The approach makes it clear that reducing the area of an ellipse at the same rate as the area of the rectangles will remove this issue. The next result confirms this intuition.

Lemma 7. *The Metropolis–Hastings algorithm with proposal Q_P produces a geometrically ergodic Markov chain when $\pi(x) = \square(x)$, from π -almost any starting point.*

Proof: We can take as a small set $C = \{y \in \mathbb{R}^2; 1 \leq y_i \leq 2\}$, i.e. the largest rectangle on the contour plot, and the Lyapunov function $V(x) = |x_2| + 1 \vee |x_1|$. For $x, y \in R$, $V(y) < V(x)$ iff $y_2 < x_2$. Note also that $\alpha(x, y) = 1$ for any $x, y \in R \cap \{y \in \mathcal{X} : y_2 < x_2\}$. It suffices, with these choices, to show that the overlap on the contour plot between the lower hemisphere of each E_x and R is larger than that between R and the upper hemisphere for any $x \in R \setminus C$, which is clearly true from inspecting the figures in Appendix C. \blacksquare

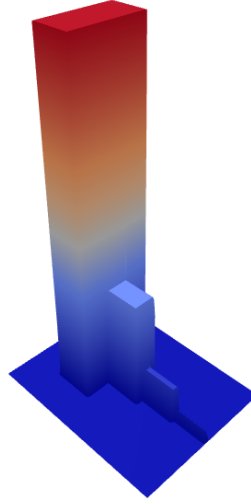


Figure 3: The rectangle density.

6. Discussion

In this paper we have analysed the ergodic behaviour of a Metropolis–Hastings method with proposal kernel $Q(x, \cdot) = \mathcal{N}(x, hG^{-1}(x))$. In one dimension we have characterised the behaviour in terms of growth conditions on $G^{-1}(x)$ and tail conditions on the target distribution, and some cases in higher dimensions have also been discussed. The fundamental question of interest was whether generalising an existing Metropolis–Hastings method by allowing the proposal covariance to change with position can alter the ergodicity properties of the sampler. We can confirm that this is indeed possible, either for the better or worse, depending on the choice of covariance. The take home points for practitioners are i) lack of sufficient care in the design of $G^{-1}(x)$ can have severe consequences (as in Lemma 4), and ii) careful choice of $G^{-1}(x)$ can have much more beneficial ones, particularly in higher dimensions, as evidenced by the ‘rectangle’ density example in Section 5.

We feel that such results can also offer insight into similar generalisations of different Metropolis–Hastings algorithms (e.g. [13, 34]). For example, it seems intuitive that any method in which the variance grows at a faster than quadratic rate in the tails is unlikely to produce a geometrically ergodic chain. There are connections between the PDRWM and some extensions of the Metropolis-adjusted Langevin algorithm [34], the ergodicity properties of which are discussed in [35]. The key difference between the schemes is the inclusion of the drift term $G^{-1}(x)\nabla \log \pi(x)/2$ in the latter. It is this term which in the main governs the behaviour of the sampler, which is why the behaviour of the PDRWM is different to this scheme (note that gradients are required for all variants, unlike in the PDRWM).

We can apply the general results to the specific variants discussed in Section 3. Provided sensible choices of regions/weights, and diminishing adaptation schemes are chosen, the Regional adaptive Metropolis–Hastings, Locally weighted Metropolis and Kernel-adaptive Metropolis–Hastings samplers should all satisfy $G^{-1}(x) \rightarrow \Sigma$ as $|x| \rightarrow \infty$, meaning they will inherit the ergodicity properties of the standard RWM (the behaviour in the centre of the space, however, will likely be different). In the State-dependent Metropolis method provided $b \leq 2$ (with suitable tuning in the equality case) then the sampler should also behave reasonably. Whether or not a large enough value of b would be found by a particular adaptation rule in the subexponential case is not entirely clear, and this could be an interesting direction of further study. The Tempered Langevin diffusion scheme, however, will fail to produce a geometrically ergodic Markov chain whenever the tails of $\pi(x)$ are lighter than that of a Cauchy distribution. In the case of Gaussian tails, for example, $G^{-1}(x) = e^{x^2/2}I$. To allow reasonable tail exploration, two pragmatic options would be to upper bound $G^{-1}(x)$ manually or use this scheme in conjunction with another, as there is evidence that the sampler can perform favourably when exploring the centre of a distribution [8]. None of the specific variants discussed here are able to mimic the local curvature of the $\pi(x)$ in the tails, so as to enjoy the favourable behaviour exemplified in Lemma 7. This is possible using Hessian information

as in [13], though should also be possible in cases where this isn't available using appropriate surrogates, at least in some cases.

It is reasonable to ask whether exploring the tails of a distribution adequately is always necessary. If the functions a practitioner is interested in estimating are such that $\int_C f(x)\tilde{\pi}(dx) \approx \int f(x)\pi(dx)$, where $\tilde{\pi}(\cdot)$ is the target restricted to the centre of the space C , then perhaps this is not so important. Some results in this direction are given in [36]. If this approach is taken, however, whether or not a sampler will perform appropriately becomes a considerably more problem-dependent question. Geometric ergodicity, whilst by no means guaranteeing sensible estimators in the non-asymptotic context, does give steps towards this in some *generality*, through (2). As mentioned earlier, it also appears to have other favourable consequences [16, 21]. As such, we feel it is a property worth establishing.

Acknowledgements

I would like to thank Alexandros Beskos, Krzysztof Łatuszyński and Gareth Roberts for several useful discussions, Michael Betancourt for proofreading the paper, and Mark Girolami for general supervision and guidance. This work was funded by a PhD scholarship from Xerox Research Centre Europe.

7. References

References

- [1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines, *The journal of chemical physics* 21 (6) (1953) 1087–1092.
- [2] W. K. Hastings, Monte carlo sampling methods using markov chains and their applications, *Biometrika* 57 (1) (1970) 97–109.
- [3] C. Sherlock, P. Fearnhead, G. O. Roberts, The random walk metropolis: linking theory and practice through a case study, *Statistical Science* (2010) 172–190.
- [4] K. L. Mengersen, R. L. Tweedie, et al., Rates of convergence of the hastings and metropolis algorithms, *The Annals of Statistics* 24 (1) (1996) 101–121.
- [5] G. O. Roberts, R. L. Tweedie, Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms, *Biometrika* 83 (1) (1996) 95–110.
- [6] S. F. Jarner, G. O. Roberts, Convergence of heavy-tailed monte carlo markov chain algorithms, *Scandinavian Journal of Statistics* 34 (4) (2007) 781–815.

- [7] G. O. Roberts, J. S. Rosenthal, Examples of adaptive mcmc, *Journal of Computational and Graphical Statistics* 18 (2) (2009) 349–367.
- [8] G. O. Roberts, O. Stramer, Langevin diffusions and metropolis–hastings algorithms, *Methodology and computing in applied probability* 4 (4) (2002) 337–357.
- [9] D. Sejdinovic, H. Strathmann, M. Lomeli Garcia, C. Andrieu, A. Gretton, Kernel adaptive metropolis–hastings, in: *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- [10] C. Andrieu, J. Thoms, A tutorial on adaptive mcmc, *Statistics and Computing* 18 (4) (2008) 343–373.
- [11] R. V. Craiu, J. Rosenthal, C. Yang, Learn from thy neighbor: Parallel-chain and regional adaptive mcmc, *Journal of the American Statistical Association* 104 (488) (2009) 1454–1466.
- [12] D. Rudolf, B. Sprungk, On a generalization of the preconditioned crank-nicolson metropolis algorithm, *arXiv preprint arXiv:1504.03461*.
- [13] M. Girolami, B. Calderhead, Riemann manifold langevin and hamiltonian monte carlo methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (2) (2011) 123–214.
- [14] S. Brooks, A. Gelman, G. Jones, X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, CRC press, 2011.
- [15] G. O. Roberts, J. S. Rosenthal, et al., General state space markov chains and mcmc algorithms, *Probability Surveys* 1 (2004) 20–71.
- [16] K. Łatuszyński, B. Miasojedow, W. Niemiro, et al., Nonasymptotic bounds on the estimation error of mcmc algorithms, *Bernoulli* 19 (5A) (2013) 2033–2066.
- [17] D. Rudolf, Explicit error bounds for markov chain monte carlo, *arXiv preprint arXiv:1108.3201*.
- [18] B. Miasojedow, Hoeffdings inequalities for geometrically ergodic markov chains on general state space, *Statistics & Probability Letters* 87 (2014) 115–120.
- [19] G. O. Roberts, J. S. Rosenthal, Geometric ergodicity and hybrid markov chains, *Electron. Comm. Probab* 2 (2) (1997) 13–25.
- [20] P. Alquier, N. Friel, R. Everitt, A. Boland, Noisy monte carlo: convergence of markov chains with approximate transition kernels, *Statistics and Computing* (2014) 1–19.
- [21] F. J. Medina-Aguayo, A. Lee, G. O. Roberts, Stability of noisy metropolis–hastings, *arXiv preprint arXiv:1503.07066*.

- [22] D. Rudolf, N. Schweizer, Perturbation theory for markov chains via wasserstein distance, arXiv preprint arXiv:1503.04123.
- [23] S. P. Meyn, R. L. Tweedie, Markov chains and stochastic stability, Cambridge university press, 2009.
- [24] G. L. Jones, J. P. Hobert, Honest exploration of intractable probability distributions via markov chain monte carlo, Statistical Science (2001) 312–334.
- [25] L. Tierney, Markov chains for exploring posterior distributions, the Annals of Statistics (1994) 1701–1728.
- [26] J. Bierkens, Non-reversible metropolis-hastings, arXiv preprint arXiv:1401.8087.
- [27] S. F. Jarner, R. L. Tweedie, et al., Necessary conditions for geometric and polynomial ergodicity of random-walk-type, Bernoulli 9 (4) (2003) 559–578.
- [28] G. O. Roberts, J. S. Rosenthal, et al., Optimal scaling for various metropolis-hastings algorithms, Statistical science 16 (4) (2001) 351–367.
- [29] S. Livingstone, M. Girolami, Information-geometric markov chain monte carlo methods using diffusions, Entropy 16 (6) (2014) 3074–3102.
- [30] C. Andrieu, É. Moulines, et al., On the ergodicity properties of some adaptive mcmc algorithms, The Annals of Applied Probability 16 (3) (2006) 1462–1505.
- [31] M. Betancourt, A general metric for riemannian manifold hamiltonian monte carlo, in: Geometric science of information, Springer, 2013, pp. 327–334.
- [32] J. D. Cook, Upper and lower bounds on the normal distribution function, available at: <http://www.johndcook.com/normalbounds.pdf>. Accessed: 2015-06-29 (October 2009).
- [33] R. M. Neal, Slice sampling, Annals of statistics (2003) 705–741.
- [34] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, M. Girolami, Langevin diffusions and the metropolis-adjusted langevin algorithm, Statistics & Probability Letters 91 (2014) 14–19.
- [35] K. Łatuszyński, G. O. Roberts, A. Thiery, K. Wolny, Discussion on ‘riemann manifold langevin and hamiltonian monte carlo methods’ (by girolami, m. and calderhead, b.), Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (2) (2011) 188–189.
- [36] N. Bou-Rabee, M. Hairer, Nonasymptotic mixing of the mala algorithm, IMA Journal of Numerical Analysis (2012) drs003.
- [37] N. L. Johnson, S. Kotz, Distributions in Statistics: Continuous Univariate Distributions: Vol.: 1, Houghton Mifflin, 1970.

Appendix A. Proofs

Appendix A.1. Proof of Lemma 2

For the log-concave case, take $V(x) = e^{s|x|}$ for some $s > 0$, and let B_A denote the integral (8) over the set A . We first break up \mathcal{X} into $(-\infty, 0] \cup (0, x - cx^{\frac{\gamma}{2}}] \cup (x - cx^{\frac{\gamma}{2}}, x + cx^{\frac{\gamma}{2}}] \cup (cx^{\frac{\gamma}{2}}, x + cx^\gamma] \cup (x + cx^\gamma, \infty)$, and show that the integral is strictly negative on at least one of these sets, and can be made arbitrarily small as $x \rightarrow \infty$ on all others. The $-\infty$ case is analogous from the tail conditions on $\pi(x)$.

On $(\infty, 0]$, we have

$$\begin{aligned} B_{(\infty, 0]} &= e^{-sx} \int_{-\infty}^0 e^{s|y|} \alpha(x, y) Q(x, dy) - \int_{-\infty}^0 \alpha(x, y) Q(x, dy), \\ &\leq e^{-sx} \int_0^\infty e^{sy} Q(-x, dy). \end{aligned}$$

The integral is now proportional to the moment generating function of a truncated Gaussian distribution (see Appendix B), so is given by

$$e^{-sx+x^\gamma h s^2/2} \left[1 - \Phi \left(x^{1-\gamma/2}/h^{1/2} - h^{1/2} s x^{\gamma/2} \right) \right].$$

A simple bound on the error function is $\sqrt{2\pi}x\Phi^c(x) < e^{-x^2/2}$ [32], so setting $\eta = x^{1-\gamma/2}/h^{1/2} - h^{1/2} s x^{\gamma/2}$ we have

$$\begin{aligned} B_{(\infty, 0]} &\leq \frac{1}{\sqrt{2\pi}} \exp \left(-2sx + \frac{hs^2}{2} x^\gamma - \frac{1}{2} (h^{-1} x^{2-\gamma} - 2sx + hs^2 x^\gamma) + \log \eta \right), \\ &= \frac{1}{\sqrt{2\pi}} \exp \left(-sx - \frac{1}{2h} x^{2-\gamma} + \log \eta \right). \end{aligned}$$

which $\rightarrow 0$ as $x \rightarrow \infty$, so we can make this arbitrarily small.

On $(0, x - cx^{\gamma/2}]$, note that $e^{s(|y|-|x|)} - 1$ is clearly negative throughout this region. So the integral is straightforwardly bounded as $B_{(0, x - cx^{\gamma/2}]} \leq 0$ for all $x \in \mathcal{X}$.

On $(x - cx^{\gamma/2}, x + cx^{\gamma/2}]$, provided $x - cx^{\gamma/2}$ is large enough that we are in the tail region, then for any y in this region

$$\alpha(x, y) \leq \exp \left(-a(y - x) + \frac{\gamma}{2} \log \left| \frac{x}{y} \right| - \frac{1}{2h} [(x - y)^2 y^{-\gamma} - (x - y)^2 x^{-\gamma}] \right).$$

A Taylor expansion of $y^{-\gamma}$ about x gives

$$y^{-\gamma} = x^{-\gamma} - \gamma x^{-\gamma-1}(y - x) + \gamma(\gamma + 1)x^{-\gamma-2}(y - x)^2 + \dots$$

and multiplying by $(y - x)^2$ gives

$$(y - x)^2 y^{-\gamma} = \frac{(y - x)^2}{x^\gamma} - \gamma \frac{(y - x)^3}{x^{\gamma+1}} + \gamma(\gamma + 1) \frac{(y - x)^4}{x^{\gamma+2}} + \dots$$

If $|y - x| = cx^{\gamma/2}$ then this is:

$$\frac{c^2 x^\gamma}{x^\gamma} - \gamma \frac{c^3 x^{3\gamma/2}}{x^{\gamma+1}} + \gamma(\gamma+1) \frac{c^4 x^{2\gamma}}{x^{\gamma+2}} + \dots$$

As $\gamma < 2$ then $3\gamma/2 < \gamma + 1$, and similarly for successive terms, meaning each gets smaller as $|x| \rightarrow \infty$. So we have for large x and $y \in (x - cx^{\gamma/2}, x + cx^{\gamma/2})$

$$(y - x)^2 y^{-\gamma} \approx \frac{(y - x)^2}{x^\gamma} - \gamma \frac{(y - x)^3}{x^{\gamma+1}}. \quad (\text{A.1})$$

Using (A.1) gives (for large enough x)

$$\alpha(x, y) \leq \exp \left(-a(y - x) + \frac{\gamma}{2} \log \left| \frac{x}{y} \right| + \frac{1}{2h} \gamma \frac{(y - x)^3}{x^{\gamma+1}} \right)$$

So we can analyse how the acceptance rate behaves. First note that for fixed $\epsilon > 0$

$$\alpha(x, x + \epsilon) \leq \exp \left(-a\epsilon + \frac{\gamma}{2} \log \left| \frac{x}{x + \epsilon} \right| + \frac{1}{2h} \gamma \frac{\epsilon^3}{x^{\gamma+1}} \right) \rightarrow \exp(-a\epsilon).$$

Similarly we find that the $e^{-a\epsilon}$ term will dominate for any ϵ for which $\epsilon^3/x^{\gamma+1} \rightarrow 0$, i.e. any $\epsilon = o(x^{\gamma+1/3})$. If $\gamma < 2$ then $\epsilon = cx^{\gamma/2}$ satisfies this condition. So for any $y > x$ in this region we can choose an x such that

$$\alpha(x, y) \leq \exp(-a(y - x) + \delta_x),$$

where δ_x can be made arbitrarily small in this region by choosing a large enough x . For the case $y < x$ here we have (for any fixed $\epsilon > 0$)

$$\alpha(x, x - \epsilon) \leq \exp \left(a\epsilon + \frac{\gamma}{2} \log \left| \frac{x}{x - \epsilon} \right| - \frac{1}{2h} \gamma \frac{\epsilon^3}{x^{\gamma+1}} \right) \rightarrow \exp(a\epsilon).$$

So by a similar argument we have $\alpha(x, y) > 1$ here for large x , as the exponential term will dominate. Combining these results we can write

$$\begin{aligned} B_{(x-cx^{\gamma/2}, x+cx^{\gamma/2})} &= \int_0^{cx^{\gamma/2}} \left[e^{(s-a)z+\delta_z} - e^{-az+\delta_z} + e^{-sz} - 1 \right] q_x(dz), \\ &= - \int_0^{cx^{\gamma/2}} (1 - e^{-sz})(1 - e^{(s-a)z+\delta_z}) q_x(dz), \end{aligned}$$

which will be strictly negative for large enough x provided $s < a$, where $q_x(\cdot)$ denotes a zero mean Gaussian distribution with the same variance as $Q(x, \cdot)$.

On $(x + cx^{\gamma/2}, x + cx^\gamma]$ we can upper bound the acceptance rate as

$$\alpha(x, y) \leq \frac{\pi(y)}{\pi(x)} \exp \left(\frac{1}{2} \log \left| \frac{G(y)}{G(x)} \right| + \frac{G(x)}{2h} (x - y)^2 \right)$$

If $y \geq x$ and $x > x_0$ then we have

$$\alpha(x, y) \leq \exp \left(-a(|y| - |x|) + \frac{1}{2h} \frac{(x - y)^2}{x^\gamma} \right).$$

For $|y - x| = cx^\eta$ this becomes

$$\alpha(x, y) \leq \exp \left(-acx^\eta + \frac{c^2}{2h} x^{2\eta - \gamma} \right)$$

So provided $\gamma > \eta$ the e^{-a} term will dominate for large x . In the equality case we have

$$\alpha(x, y) \leq \exp \left(\left(\frac{c^2}{2h} - a \right) cx^\gamma \right),$$

so provided we choose c such that $a > c^2/2h$ then the acceptance rate will also decay exponentially. Because of this we have

$$\begin{aligned} B_{(x+cx^{\gamma/2}, x+cx^\gamma]} &\leq \int_{A_4} e^{s(y-x)} \alpha(x, y) Q(x, dy), \\ &\leq e^{(c^2/2h+s-a)cx^{\gamma/2}} Q(x, (x+cx^{\gamma/2}, x+cx^\gamma]), \end{aligned}$$

so provided $a > c^2/2h + s$ then this term can be made arbitrarily small.

On $(x+cx^\gamma, \infty)$ using the same properties of truncated Gaussians and error function bounds we have

$$\begin{aligned} B_{(x+cx^\gamma, \infty)} &\leq e^{-sx} \int_{x+cx^\gamma}^{\infty} e^{sy} Q(x, dy), \\ &= e^{s^2x^{\gamma/2}} \Phi^c((c-s)x^\gamma) \leq \exp \left(\frac{-c(c-2s)}{2} x^\gamma \right), \end{aligned}$$

which can be made arbitrarily small provided $c > 2s$. ■

For the subexponential case, the proof is similar. Take $V(x) = e^{s|x|^\beta}$, and divide \mathcal{X} up into the same regions. Outside of $(x - x^{\gamma/2}, x + x^{\gamma/2}]$ the same arguments show that the integral can be made arbitrarily small. On this set, note that in the tails.

$$(x + cx^\eta)^\beta - x^\beta = \beta x^{\eta+\beta-1} + \beta(\beta-1)x^{2\eta+\beta-2} + \dots$$

For $y - x = cx^\eta$, then for $\eta < 1 - \beta$ this becomes negligible, otherwise it will grow as x does. So in this case we further divide the typical set into $(x, x + cx^{1-\beta}] \cup (x + cx^{1-\beta}, x + cx^{\gamma/2}]$. On $(x - cx^{1-\beta}, x + cx^{1-\beta})$ the integral is bounded above by $e^{-c_1} Q(x, (x - cx^{1-\beta}, x + cx^{1-\beta})) \rightarrow 0$, for some suitably chosen $c_1 > 0$. On $(x - cx^{\gamma/2}, x - cx^{1-\beta}] \cup (x + cx^{1-\beta}, x + cx^{\gamma/2}]$ then for $y > x$ we have $\alpha(x, y) \leq e^{-c_2(y^\beta - x^\beta)}$, so we can use the same argument as in the log-concave case to show that the integral will be strictly negative in the limit. ■

Appendix A.2. Proof of Lemma 3

Here a typical proposal will be $y = x \pm \xi\sqrt{h}x$ for x sufficiently large, meaning $|x - y| = \xi\sqrt{h}x$, with $\xi \sim \mathcal{N}(0, 1^2)$. For now we assume both x and y are in the tail regime, meaning $G(y) \propto y^{-2}$ and similarly for $G(x)$ (we make this concrete later). We can also take $\pi(y)/\pi(x) = x^p/y^p$ here.

For $y = (1 + \xi\sqrt{h})x$ then in the tails the acceptance rate becomes

$$\alpha(x, y) = 1 \wedge \frac{1}{(1 + \xi\sqrt{h})^{p+1}} \exp\left(\frac{\xi^3\sqrt{h}}{2} \left[\frac{2 + \xi\sqrt{h}}{(1 + \xi\sqrt{h})^2}\right]\right),$$

which is completely independent of x .

Take $V(x) = 1 \vee |x|^s$, for some $s < 1$ which is suitably small that $\int V(y)\pi(dy) < \infty$, together with an extra restriction which we specify later. Then $V(y)/V(x)$ becomes independent of x also. The integral of interest can now be re-written in terms of ξ , with $m(\cdot)$ a standard Gaussian measure, $\phi(\xi)$ its density, and $\alpha_h(\xi)$ the acceptance rate. So in most of the regions we consider we can choose x large enough that the integral in question is

$$\int \left[|1 + \xi\sqrt{h}|^s - 1\right] \alpha_h(\xi) m(d\xi). \quad (\text{A.2})$$

We therefore need to show that this integral is strictly negative for h small enough, and take care of the values of y which may not fall into this region.

Again denoting (8) integrated over a region A as B_A , we can break (A.2) up into

$$\begin{aligned} B_{(\infty, \infty)} &= B_{(-\infty, -2h^{-\frac{1}{2}})} + B_{(-2h^{-\frac{1}{2}}, -\delta h^{-\frac{1}{4}})} + B_{(-\delta h^{-\frac{1}{4}}, \delta h^{-\frac{1}{4}})} + B_{(\delta h^{-\frac{1}{4}}, \infty)}, \\ &= B_{H_1} + B_{H_2} + B_{H_3} + B_{H_4}. \end{aligned}$$

It is clear that all of these integrals can be made arbitrarily close to zero by making h small enough. The goal is to show that $B_{(\infty, \infty)} < 0$ for all $h \in (0, h_0)$. We proceed by finding the order of h of each B_{H_i} .

On $H_1 = (-\infty, -2h^{-\frac{1}{2}})$ we have

$$B_{H_1} \leq \frac{1}{\sqrt{2\pi}} \int_{H_1} \left[|1 + \xi\sqrt{h}|^s - 1\right] \exp\left(-\frac{\xi^2}{2}\right) d\xi$$

Use the change of variables $\gamma = 1 + \xi\sqrt{h}$ gives

$$B_{H_1} \leq \int_{-\infty}^{-1} [|\gamma|^s - 1] m(d\gamma) = \int_1^\infty (\eta^s - 1) m(d\eta) < \int_1^\infty \eta m(d\eta),$$

with $\eta \sim \mathcal{N}(-1, h)$, as $s < 1$. Using results for truncated Gaussians, we have

$$\begin{aligned} \int_1^\infty \eta m(d\eta) &= -\Phi\left(-\frac{2}{\sqrt{h}}\right) + \sqrt{h}\phi\left(\frac{2}{\sqrt{h}}\right) \frac{\Phi\left(-\frac{2}{\sqrt{h}}\right)}{1 - \Phi\left(\frac{2}{\sqrt{h}}\right)}, \\ &= -\Phi^c\left(\frac{2}{\sqrt{h}}\right) + \sqrt{h}\phi\left(\frac{2}{\sqrt{h}}\right). \end{aligned}$$

The lower bound on Φ^c from [32] gives

$$B_{H_1} \leq \frac{2+h}{4+h} \sqrt{\frac{h}{2\pi}} \exp\left(-\frac{2}{h}\right).$$

On $H_2 = (-2h^{-\frac{1}{2}}, -\delta h^{-\frac{1}{4}})$, the function $[|1 + \xi\sqrt{h}|^s - 1]$ is negative in H_2 , so this integral is trivially bounded as ≤ 0 for any h . Note that this is the entire set of y 's for which (A.2) is not the correct integral.

On $H_3 = (-\delta h^{-1/4}, \delta h^{-1/4})$ recall that the acceptance probability is

$$\alpha_h(\xi) = \exp\left(-(p+1)\log(1 + \xi\sqrt{h}) + \frac{\xi^3 h}{2} \left[\frac{2 + \xi\sqrt{h}}{(1 + \xi\sqrt{h})^2} \right]\right)$$

For any $\xi > 0$ we have

$$\frac{2 + \xi\sqrt{h}}{(1 + \xi\sqrt{h})^2} < \frac{2(1 + \xi\sqrt{h})}{(1 + \xi\sqrt{h})^2} < 2, \quad \text{so} \quad \frac{\xi^3 h}{2} \left[\frac{2 + \xi\sqrt{h}}{(1 + \xi\sqrt{h})^2} \right] < \xi^3 h,$$

meaning

$$\alpha_h(\xi) < \exp\left(-(p+1)\log(1 + \xi\sqrt{h}) + \xi^3 h\right).$$

We would like to write this as $(1 + \xi\sqrt{h})^{-a}$ for some $a > 0$. If $\delta h^{\frac{1}{4}} < 1$ we can use a Taylor expansion with remainder $\log(1 + x) = x - x^2/2 + r^3/3$ for some $r \in (0, x)$ to get the bound $x - x^2/2 \leq \log(1 + x)$ for $0 \leq x < 1$. For any $b < p+1$ then

$$b \log(1 + \xi\sqrt{h}) > b \left(\xi\sqrt{h} - \frac{\xi^2 h}{2} \right) > \frac{b\xi\sqrt{h}}{2} > \xi^3 h \quad \text{for } \xi \in (0, \delta h^{-\frac{1}{4}}), \delta < \sqrt{\frac{b}{2}}.$$

So provided δ is chosen in this way then $\exists a > 0$ such that $\alpha_h(\xi) \leq (1 + \xi\sqrt{h})^{-a}$ for $\xi \in (0, \delta h^{-\frac{1}{4}})$ and $\alpha = 1$ for $\xi \in (-\delta h^{-\frac{1}{4}}, 0)$ (by simply reversing the signs in the above inequalities). Now the integral of interest can be written

$$B_{H_3} \leq \int_0^{\delta h^{-\frac{1}{4}}} \left[(1 + \xi\sqrt{h})^{(s-a)} - (1 + \xi\sqrt{h})^{-a} + (1 - \xi\sqrt{h})^s - 1 \right] m(d\xi).$$

So we need to bound

$$\int (1 + \xi\sqrt{h})^{s-a} m(d\xi) - \int (1 + \xi\sqrt{h})^{-a} m(d\xi) + \int (1 - \xi\sqrt{h})^s m(d\xi) - \frac{1}{2} \Phi(\delta h^{-\frac{1}{4}}).$$

Upper and lower bounds for $g(\xi) = (1 + \xi\sqrt{h})^{-a}$ on $(0, \delta h^{-\frac{1}{4}})$ are

$$\begin{aligned} g_u(\xi) &= m_u(a)\xi + 1, \quad m_u(a) = \frac{h^{\frac{1}{4}}}{\delta} \left[(1 + \delta h^{\frac{1}{4}})^{-a} - 1 \right], \\ g_l(\xi) &= m_l(a)\xi + 1, \quad m_l(a) = -a\sqrt{h}. \end{aligned}$$

The first is a straight line through $g(\delta h^{-\frac{1}{4}})$ and $g(0) = 1$, the second is the straight line through $g(0) = 1$ with gradient $g'(0)$ (as the function is concave). This gives upper and lower bounds for the first two integrals as

$$m_u(a-s)\Psi_h + \Phi(\delta h^{-\frac{1}{4}}) - \frac{1}{2}, \quad \text{and} \quad m_l(a)\Psi_h + \Phi(-\delta h^{\frac{1}{4}}) - \frac{1}{2}.$$

where $\Psi_h = \phi(\delta h^{-\frac{1}{4}}) - 1/\sqrt{2\pi} < 0$. We can construct a similar Taylor Series upper bound for $(1 - \xi\sqrt{h})^s$ as a straight line with gradient $m_u^* = -s\sqrt{h}$ (as this function is concave), meaning the total bound of interest is

$$\begin{aligned} B_{H_3} &\leq (m_u(a-s) - m_l(a) + m_u^*)\Psi_h, \\ &= \left((a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta} \left((1 + \delta h^{\frac{1}{4}})^{s-a} - 1 \right) \right) \Psi_h, \\ &= C_{H_3} \exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) - C_{H_3}, \end{aligned}$$

where $C_{H_3} = (a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta} \left((1 + \delta h^{\frac{1}{4}})^{s-a} - 1 \right)$. It is clear that C_{H_3} is positive, as

$$1 + (a-s)\delta h^{\frac{1}{4}} > 1 > \left(1 + \delta h^{\frac{1}{4}}\right)^{-(a-s)}$$

because $(a-s) > 0$.

On $H_4 = (\delta h^{-1/4}, \infty)$, bounding in the same way as for H_1 , we set $\gamma = 1 + \xi\sqrt{h}$, meaning $\gamma \sim \mathcal{N}(1, h)$. Then

$$B_{H_4} \leq \int_{\delta h^{-\frac{1}{4}}}^{\infty} [|\gamma|^s - 1] m(d\gamma),$$

which can be re-written

$$\begin{aligned} \mathbb{E}_{\varpi} [|\gamma|^s - 1] \Phi^c(\delta h^{-\frac{1}{4}}) &\leq \mathbb{E}_{\varpi} [\gamma] \Phi^c(\delta h^{-\frac{1}{4}}), \\ &= (1 + \delta h^{\frac{1}{4}}) \Phi^c(\delta h^{-\frac{1}{4}}) + \sqrt{h} \phi(\delta h^{-\frac{1}{4}}), \end{aligned}$$

where ϖ is now a truncated Gaussian distribution on $(1 + \delta h^{\frac{1}{4}}, \infty)$ with mean 1 and variance h . Using the upper bound on Φ^c gives

$$\begin{aligned} B_{H_4} &\leq (1 + \delta h^{\frac{1}{4}}) \frac{1}{\sqrt{2\pi}} \frac{h^{\frac{1}{4}}}{\delta} \exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) + \sqrt{\frac{h}{2\pi}} \exp\left(-\frac{\delta^2}{2\sqrt{h}}\right), \\ &= \sqrt{\frac{h^{\frac{1}{4}}}{2\pi}} \left(2h^{\frac{1}{4}} + \frac{1}{\delta}\right) \exp\left(-\frac{\delta^2}{2\sqrt{h}}\right), \\ &= C_{H_4} \exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) \end{aligned}$$

Combining inequalities, we can get a very loose upper bound on the integral as

$$B_{(-\infty, \infty)} \leq (C_{H_4} + C_{H_3}) \exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) + C_{H_1} \exp\left(-\frac{2}{h}\right) - C_{H_3}.$$

The exponentials are the dominant terms in the first two expressions, as they shrink to zero much faster than any of the C_{H_i} terms (which still depend on h). To see that this is the case for C_{H_3} , note that $(1 + \delta h^{\frac{1}{4}})^{s-a} = 1 + (s-a)\delta h^{\frac{1}{4}} + O(h^{\frac{1}{2}})$, so that

$$\begin{aligned} C_{H_3} &= (a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta} \left((1 + \delta h^{\frac{1}{4}})^{s-a} - 1 \right), \\ &= (a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta} \left(-(a-s)\delta h^{\frac{1}{4}} + O(h^{\frac{1}{2}}) \right), \\ &= O(h^{\frac{3}{4}}). \end{aligned}$$

It is more straightforward to see that C_{H_1} and C_{H_4} are both $O(h^{\frac{1}{2}})$. Because of this, we can always choose a h small enough that the last term is arbitrarily larger than all others in the expression, meaning that the integral is strictly negative, as required. ■

Appendix A.3. Proof of Lemma 4

The goal is to show

$$\limsup \int \alpha(x, y) Q(x, dy) = 0.$$

The general strategy will be to find some set

$$A_{x,\epsilon} := \{y \in \mathcal{X} : \alpha(x, y) \geq \epsilon\}.$$

In words, a set which shows the potential candidate moves which have a non-negligible probability of acceptance. We will then establish that $Q(x, A_{x,\epsilon}) \rightarrow 0$ as $x \rightarrow \infty$, for any $\epsilon > 0$.

First recall that for the algorithm in general the acceptance probability for a proposal y is

$$\alpha(x, y) = \frac{\pi(y)|G(y)|^{\frac{1}{2}}}{\pi(x)|G(x)|^{\frac{1}{2}}} \exp \left(-\frac{1}{2h}(y-x)^2[G(y) - G(x)] \right).$$

If $G(x) = O(|x|^{-\gamma})$, then for large enough x and y the acceptance probability is

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \left(\frac{|x|}{|y|} \right)^{\frac{\gamma}{2}} \exp \left(-\frac{c}{2h}(x-y)^2 \left[\frac{1}{|y|^\gamma} - \frac{1}{|x|^\gamma} \right] \right).$$

As each $Q(x, \cdot)$ is a Gaussian distribution, we consider a ‘typical set’ to be

$$T_x = \left(x - 2\sqrt{h}x^{\gamma/2}, x + 2\sqrt{h}x^{\gamma/2} \right).$$

For any x , $Q(x, T_x) \approx 0.96$. If we can show that i) for large enough x , $A_{x,\epsilon} \subset T_x$, and ii) the ratio $Q(x, A_{x,\epsilon})/Q(x, T_x) \rightarrow 0$ then we will have established the result.

First we note that for $|y|$ larger than x then if $\pi(x) \in C_0(\mathbb{R})$ then in the tails $\pi(y)/\pi(x) \leq 1$, so we can say

$$\alpha(x, y) \leq \left(\frac{x}{|y|} \right)^{\frac{\gamma}{2}} \exp \left(-\frac{c}{2h} \left[\frac{(x-y)^2}{|y|^\gamma} - \frac{(x-y)^2}{x^\gamma} \right] \right).$$

Since if $y = x$ then $\alpha(x, y) = 1$, we will only concern ourselves with $|y| > |x|$. In effect we are now considering the set $A_{x,\epsilon} \cup (-x, x)$, but since this is strictly larger than $A_{x,\epsilon}$ it will give us the result. For $y > x$, if we write $y = x + z$ for some $z > 0$ (and do similar in the other tail), we can see that

$$\alpha(x, x+z) \leq \left(\frac{x}{x+z} \right)^{\frac{\gamma}{2}} \exp \left(-\frac{cz^2}{2h(x+z)^\gamma} + \frac{cz^2}{2hx^\gamma} \right).$$

As $x \rightarrow \infty$, the first term on the right-hand side will tend to something greater than zero for $z = O(x)$ and decay to zero for the set of z 's that grow at a larger rate than x . Inside the exponential, the term $cz^2/2h(x+z)^\gamma \rightarrow 0$ for any z as x grows. The last term $cz^2/2hx^\gamma$ will only increase with x for the set of z 's that grow at a faster rate than $x^{\gamma/2}$. If we denote this set of 'extreme' values for y which would be accepted as $E_{x,\epsilon} = A_{x,\epsilon} \cap T_x^c$, then it is clear that $Q(x, E_{x,\epsilon}) \rightarrow 0$ for any $\epsilon > 0$, as $E_{x,\epsilon} \sim (-\infty, -x^{\gamma/2+\delta}) \cup (x^{\gamma/2+\delta}, \infty)$ for some $\delta > 0$, and this set will be sent deeper and deeper into the tails of $Q(x, \cdot)$ as $|x|$ grows.

So now we can focus on $A_{x,\epsilon} \cap T_x$, or equivalently consider the set of possible z values in $(-2x^{\gamma/2}, 0) \cup (0, 2x^{\gamma/2})$. For any of these the dominant term in $\alpha(x, x+z)$ will be $(x/(x+z))^{\gamma/2}$, so the acceptance rate will be strictly decreasing in z on this set. Hence we need only examine the boundary points, $y = x + 2\sqrt{h}x^{\gamma/2}$ and $y = x - 2\sqrt{h}x^{\gamma/2}$, and show that these both decay to zero as $x \rightarrow \infty$.

For $y = x + 2\sqrt{h}x^{\gamma/2}$ the acceptance rate becomes

$$\begin{aligned} \alpha(x, y) &\leq \left(\frac{x}{x + 2\sqrt{h}x^{\gamma/2}} \right)^{\gamma/2} \exp \left(-\frac{c}{2h} \left[\frac{4\sqrt{h}x^\gamma}{|x + 2\sqrt{h}x^{\gamma/2}|^\gamma} - 4\sqrt{h} \right] \right), \\ &\leq \left(\frac{x}{x + 2\sqrt{h}x^{\gamma/2}} \right)^{\gamma/2} \exp \left(\frac{2c}{\sqrt{h}} \right), \\ &\rightarrow 0. \end{aligned}$$

And for $y = x - 2\sqrt{h}x^{\gamma/2}$, noting that for large x $|x - 2\sqrt{h}x^{\gamma/2}| > \sqrt{h}x^{\gamma/2}$, we have

$$\begin{aligned} \alpha(x, y) &\leq \left(\frac{x}{\sqrt{h}x^{\gamma/2}} \right)^{\gamma/2} \exp \left(\frac{2c}{\sqrt{h}} \right) \exp \left(-\frac{c}{2h} \left[\frac{4\sqrt{h}x^\gamma}{x^{\gamma^2/2}} \right] \right), \\ &\leq \left(\frac{x}{\sqrt{h}x^{\gamma/2}} \right)^{\gamma/2} \exp \left(\frac{2c}{\sqrt{h}} \right), \\ &\rightarrow 0. \end{aligned}$$

■

Appendix B. Needed facts about truncated Gaussian distributions

Here we collect some elementary facts used in the article. For more detail see e.g. [37]. If X follows a truncated Gaussian distribution $\mathcal{N}_{[a,b]}^T(\mu, \sigma^2)$ then it has density

$$f(x) = \frac{1}{\sigma Z_{a,b}} \phi\left(\frac{x - \mu}{\sigma}\right) \mathbb{1}_{[a,b]}(x),$$

where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, $\Phi(x) = \int_{-\infty}^x \phi(y)dy$ and $Z_{a,b} = \Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma)$. Defining $B = (b - \mu)/\sigma$ and $A = (a - \mu)/\sigma$, we have

$$\mathbb{E}[X] = \mu + \frac{\phi(A) - \phi(B)}{Z_{a,b}} \sigma$$

and

$$\mathbb{E}[e^{tX}] = e^{\mu t + \sigma^2 t^2/2} \left[\frac{\Phi(B - \sigma t) - \Phi(A - \sigma t)}{Z_{a,b}} \right].$$

In the special case $b = \infty$, $a = 0$ this becomes $e^{\mu t + \sigma^2 t^2/2} \Phi(\sigma t)/Z_{a,b}$.

Appendix C. Rectangle contour plots

The contour plots show the region of proposals which would be accepted if the current point is given by the green dot. The area in the lower half of the ellipse which is coloured yellow is larger than that in the upper half (shown in red), implying that on average the vertical coordinate (and hence $V(x)$) will be smaller for the next point in the chain.

